

2020 Kaggle DS & ML Survey

- Methodology and survey flow logic

Survey Methodology: **2**

Survey Flow Logic: **3**

Survey Methodology:

- The 2020 Kaggle DS & ML Survey received 20,036 usable responses from participants in 171 different countries and territories. If a country or territory received less than 50 respondents, we grouped them into a group named “Other” for anonymity.
- An invitation to participate in the survey was sent to the entire Kaggle community (anyone opted-in to the Kaggle Email List). The survey was also promoted on the Kaggle website and on the Kaggle Twitter channel.
- The survey was live from 10/07/2020 to 10/30/2020. We allowed respondents to complete the survey at any time during that window.
- Responses to multiple choice questions (only a single choice can be selected) were recorded in individual columns. Responses to multiple selection questions (multiple choices can be selected) were split into multiple columns (with one column per answer choice).
- To protect the respondents’ privacy, free-form text responses were not included in the public survey dataset, and the order of the rows was shuffled (responses are not displayed in chronological order).
- We excluded respondents that were flagged by our survey system as “Spam”.
- The survey data was released under a CC 2.0 license:
<https://creativecommons.org/licenses/by/2.0/>

Survey Flow Logic:

- The full list of questions and answer choices can be found in the file: `kaggle_survey_2020_answer_choices.pdf`. This file contains footnotes that describe which questions were asked to which respondents. Additional details are described below.
- Respondents with the most experience were asked the most questions. For example, students and unemployed persons were not asked questions about their employer. Likewise, respondents that do not write code were not asked questions about writing code.
- For questions about cloud computing products, students and respondents that have never spent money in the cloud were given an alternate set of questions that asked them “what products they would like to become familiar with” instead of asking them “which products they use most often”. For questions with alternative phrasing, the questions were kept separate, and question types were labeled with either an “A” or a “B” (e.g. Q25A, Q25B, ... , Q34A, Q34B).
- Follow-up questions were only asked to respondents that answered the setup question affirmatively.
 - Question 18 and Question 19 (which specific ML methods) were only asked to respondents that selected the relevant answer choices for Question 17 (which categories of algorithms).
 - Question 27-A and Question 28-A (which specific AWS/Azure/GCP products) were only asked to respondents that selected the relevant answer choices for Question 26-A (which of the following companies).
 - Question 30 (which specific product) was only asked to respondents that selected more than one choice for Question 29-A (which of the following products).
 - Question 32 (which specific product) was only asked to respondents that selected more than one choice for Question 31-A (which of the following products).
 - Question 34-A (which specific product) was only asked to respondents that answered affirmatively to Question 33-A (which of the following categories of products).